Typological evidence against universal effects of referential scales on case alignment

Bickel, Balthasar ; Witzlack-Makarevich, Alena ; Zakharko, Taras

Originally published at:
Bickel, Balthasar; Witzlack-Makarevich, Alena; Zakharko, Taras (2014). Typological evidence against universal effects of referential scales on case alignment. In: Bornkessel-Schlesewsky, Ina; Malchukov, Andrej; Richards, Marc. Scales and Hierarchies: a cross-disciplinary perspective on referential hierarchies. Berlin: De Gruyter Mouton, 7-44.

Balthasar Bickel, Alena Witzlack-Makarevich, and
Taras Zakharko

2 Typological evidence against universal effects of referential scales on case alignment*

If a language develops differential subject or differential object marking by case or adpositions, this is widely hypothesized to result from a universal effect of referential scales. The effect can be understood as a universal correlation between the odds of overt case marking and scale ranks (a negative correlation for subjects, a positive one for objects), or as an implicational universal proposing that, if a language has a split in case marking, this split fits a universal scale. We test both claims with various versions of scale definitions by statistically estimating diachronic biases towards correlations or scale-fitting in an areally stratified sample of over 460 case systems worldwide. For most scales tested, results suggest evidence against universal preferences towards universal scale effects under either a correlational or an implicational model. For binary part-of-speech and information-structure distinction and object marking, the evidence for universal effects is inconclusive. What we do find, by contrast, is highly significant area effects: case-marking splits tend to have developed and spread in Eurasia and the New-Guinea/Australia ('Sahul') macro-areas. This suggests that any replication of scale effects across language families is a side-effect of areal diffusion rather than of universal principles in grammar or cognition.

1 Introduction

Typological generalizations are often first based on small-scale surveys or contrastive analyses of a few languages, and it is typically only later, after much additional empirical groundwork, that they can be evaluated through rigorous

* This research was supported by Grant No. BI 799/3-1 from the *Deutsche Forschungsgemeinschaft*. Bickel designed the study and wrote the paper, Bickel and Zakharko performed the statistical analyses, and Witzlack-Makarevich did most of the data analysis. All computations were done in R (R Development Core Team 2012), with the added packages `vcd` (Meyer et al. 2006), `MASS` (Venables & Ripley 2002), and `glmperm` (Werft & Potter 2010). We thank an anonymous reviewer for helpful comments.

quantitative analysis. Many initial generalizations have been corroborated in this way over time (as is the case, for example, with the bulk of Greenberg's word order correlations: Dryer 1992, Cysouw 2011, Bickel 2011b), but other initial generalizations have turned out to be spurious (as is the case, for example, with claims about a principled distinction between 'agglutinating' vs. 'fusional' morphologies: Haspelmath 2009). Some initial generalizations, however, have never been subject to systematic and large-scale quantitative analysis. One such generalization is the idea that, universally, some kind of referential scale governs the kinds of case or adposition markings we find, such that, for example, first and second person pronouns stand a higher chance for accusative as opposed to ergative case marking.[1]

The idea was developed in the late 70s (Silverstein 1976, Moravcsik 1978, Comrie 1981, DeLancey 1981, among others) and despite the lack of large-scale empirical tests, it is now widely taken to be an established finding. Aissen (1999), for example, counts the idea "among the most robust generalizations in syntactic markedness" and accepts a version of the idea as reflecting an inviolable component of "universal grammar" (also cf. Kiparsky 2008).

In this paper we subject the idea of scale effects on case marking to empirical testing against data from a large typological database with world-wide coverage. In order to do so, we first discuss various versions of the idea and reformulate them as precise and testable hypotheses (Section 2). In Section 3 we introduce a method for testing these hypotheses as typological claims on how languages are expected to develop over time and we explain our data coding procedure. The results of our tests are presented in Section 4. Section 5 discusses the findings and the concluding section (Section 6) compares the findings to earlier results and suggests directions for future research.

2 Claims and hypotheses

The idea of scale effects on case alignment does not easily translate into precise and testable hypotheses because there are many ways in which the idea can be spelled out – specifically, the hypotheses can be understood as absolute universals ('laws of grammar') or as probabilistic trends ('statistical universals'); as

1 In the following we use the term 'case' as a cover term for any dependent-marking of argument roles, including adpositional marking and generalizing across the kind of morphology and phonology involved. By the same token we abstract away from the distribution of case exponents inside an NP: an NP counts as case-marked if there is some nonzero case exponence somewhere in the NP – even if this is limited to determiners, as is often the case for example in German.

affecting overt case exponence (Comrie 1981) or as affecting alignment in any kind of grammatical relation (Silverstein 1976); as predicting the type of entire alignment or marking systems or as predicting correlations of alignment or marking systems with ranks on the scale. In the following we discuss these different ways of spelling out the basic idea.

2.1 Universals, variation, and exceptions

When hypothesized universals are shown to have exceptions, there are two possible responses: one can try and 'explain away' the exceptions and thereby reduce the variation (i.e. choose a 'reductionist' approach); the hypothesized universal is then 'absolute', inviolable. Alternatively, one can measure the variation and try to explain the resulting distribution (i.e. choose a 'variationist' approach); the universal is then 'statistical' and violable to a degree that can be measured.

An example for a 'reductionist' approach is Kiparsky's (2008) tentative analysis of Arrernte: in Arrernte (e.g. Mparntwe Arrernte: Wilkins 1989), the first person singular pronoun and nouns have ergative case marking, all other pronouns show accusative alignment. Under a reductionist analysis, this unexpected distribution can be accounted for by claiming that despite its appearance, the first person pronoun is a noun in this language, i.e. that it belongs to the same part of speech as lexical nouns, while other pronouns constitute a part of speech of their own. The challenge for such an approach is of course to find independent evidence for the analysis. So far, we are not aware of any such evidence although we cannot obviously exclude the possibility of finding evidence in the future. The intrinsic risk of the reductionist approach is non-testability because there is always a non-zero chance of discovering further apparent counterexamples of the Arrernte kind, and for these, we cannot anticipate whether they can be explained away.

Under a 'variationist' approach, the Arrernte distribution counts as a real exception, and the question then is how many such exceptions there are, and whether they are less frequent than distributions that match the expectations. In this paper, we follow this variationist approach exclusively. The basic hypothesis then is that there are universal principles of referential scale effects that 'push' the development of case distributions in certain ways. As a result, case distributions that fit the principles are predicted to be more common than others. The null hypothesis against which this prediction can be statistically tested, is that case distributions are not affected by universal principles of referential scale effects, but instead follow from what looks like random diachronic

fluctuation, i.e. current case distributions follow from whatever diachronies they went through. For example, if an ergative arose from an instrumental, we expect it to be limited to inanimates. This will then mimic a referential scale effect, but under the null hypothesis, it will be a mere epiphenomenon (cf. Garrett 1990). Indeed, under the null hypothesis, it will just be as likely that, for example, an ergative case system decays in lexical nouns but survives in pronouns (cf. Filimonova 2005). This will then lead to systems that do not mimic any referential scale effect and instead look like violations of such effects.

2.2 Marking, markedness, and alignment

Ever since its original formulations, the idea of scale effects has had two possible interpretations: under one interpretation (associated with Comrie 1981), referential scales affect the distribution of overt case exponence: low-ranking A arguments and high-ranking P arguments are predicted to carry overt case markers ('ergative' and 'accusative', respectively) while high-ranking A and low-ranking P arguments are predicted to carry no overt case markers (zero forms).[2] This can be extended to predictions on the phonological amount of case exponence, as proposed by Keine & Müller (2015).

An alternative interpretation (associated with Silverstein 1976), makes predictions not about overt marking patterns but about abstract markedness relations: under this interpretation, low-ranking A arguments and high-ranking P arguments are predicted to be mapped into marked grammatical relations, while high-ranking A and low-ranking P arguments are predicted to be mapped into unmarked grammatical relations. The terms 'marked' and 'unmarked' are used in a classical structuralist sense in this approach and describe which grammatical relation is structurally more constrained or specified than the other. There are many technical ways in which the relevant constraints and specifications can be spelled out, but the one that is most often associated with Silverstein's original proposal has to do with the alignment of grammatical relations, i.e. the way arguments are mapped into sets. Given this, the relevant specifications are defined by alignment sets: the sets {S,P}, {S,A} and {S,A,P} are all less specific than the sets {A} and {P}. Therefore, we expect low-ranking A arguments and high-ranking P arguments to be associated with {A} and {P} relations, respectively, while high-ranking A and low-ranking P arguments are expected to be

2 We use A and P as symbols for proto-agent and proto-patient arguments of bivalent verbs in the sense of Dowty (1991). S stands for the sole argument of monovalent predicates.

associated with the more general sets that also include S, i.e. {S,A,P} or {S,A} for high-ranking A arguments, and {S,A,P} or {S,P} for low-ranking P arguments.

Silverstein's interpretation makes predictions for any kind of alignment set, i.e. any kind of grammatical relation. This includes not only alignment sets defined by case marking but also alignment sets defined by agreement systems, conjunction reduction, or whatever syntactic structures select specific arguments to the exclusion of others. Comrie's interpretation, by contrast, is limited to case marking. Bickel (2008) and Bickel et al. (in press) demonstrate that the generalization beyond case marking has no empirical support: tested against worldwide databases on alignment splits in agreement systems, there is no trend for such systems to follow the predictions. For alignments in other syntactic structures, we lack sufficiently rich databases, but a preliminary survey reveals no systematic trend either. For diathesis in particular, Bickel & Gaenszle (2007) show that there is no systematic association of scale ranks with passivization as opposed to antipassivization: first person P arguments, for example, are required to be passivized in just as many languages as they are required to be antipassivized. For grammatical relations targeted by relative clause constructions, there are both languages where higher-ranking arguments are preferred and languages where lower-ranking arguments are preferred (Bickel 2011a).

With regard to case systems, Silverstein's and Comrie's versions make the same predictions to the extent that structurally unmarked relations tend to have less phonological exponence than structurally marked relations. Our database contains one single language that systematically deviates from this in having a morphologically marked {S,A} case, and shows at the same time an alignment split based on a referential scale: this is Middle Atlas Berber where the marked nominative (in the form of a 'construct state') is restricted to low-ranking S and A arguments. This fits Comrie's prediction that low-ranking A arguments receive morphologically overt marking. In return, it violates Silverstein's version of scale effects because low-ranking P arguments are mapped into a structurally marked grammatical relation: P is mapped into the {P} set, which is structurally marked relative to the less specific {S,A} set. However, this is one language and we cannot make any statistical inferences from this.[3]

3 The Australian language Mangarayi (Merlan 1982) is one further case of a language with a split and a marked 'nominative' (*ŋarla-* in the feminine, *ṇa-* elsewhere) in opposition to a (slightly) less marked 'accusative' (*ŋan-* in the feminine, zero elsewhere), but in this language, the referential split affects S rather than A or P: low-ranking S arguments and all P arguments are in the 'accusative', high-ranking S arguments and all A arguments are in the 'nominative'. Comrie's hypothesis makes no prediction on this. In terms of alignment, the low-ranking arguments show {S,P} alignment, while high-ranking arguments display {S,A} alignment. This fits Silverstein's predictions.

Since there is no evidence for scale effects beyond case-marking and since for all but one relevant language, structural markedness correlates with morphological markedness, we focus on case marking and use structural markedness, i.e. alignment sets, as a proxy for morphological markedness.[4]

The only problematic case for this approach is presented by double-oblique alignment {A,P} vs. {S} that contrasts with ergative or accusative alignment. An example is Vafsi, a Northwestern Iranian language. In past tense clauses of this language, nominal A arguments are in what is called the oblique case; P arguments are also in the same oblique case if they rank high in discourse status, e.g. by being definite (1a). Lower-ranking (e.g. indefinite) P arguments, by contrast, are in the 'direct' case (1b), which also covers S arguments (1c):

(1) Vafsi (ISO639.3:vaf; Northwestern Iranian; Indo-European; Stilo 2004)

a. *luás-i* *kærg-é=s* *b̲ǽ-værdæ.*
fox-OBL chicken-OBL.F=3s PUNCT-took
A P
'The fox took the chicken.'

b. *in* *luti-an* *yey xær=esan* *æ-rúttæ.*
DEM wise.guys-OBL.PL one donkey.DIR=3p DUR-sold
A P
'These wise guys were selling a donkey.'

c. *zení-e* *há-nešesd-end.*
woman-PL.DIR PVB-sat-3p
S
'The women sat down.'

Such a system sets up a contrast between {A,P} for high-ranking P arguments and {S,P} for low-ranking P arguments. Since the two alignments contain the same number of specifications (two each), one could argue that they are equally marked. However, closer inspection of the morphological markedness and of what we know from the history of these languages (Haig 2008) suggests that {A,P}, i.e. the oblique forms, represents the structurally marked forms, while {S,P}, i.e. the direct forms, represent the unmarked forms. In addition, a case

4 We do not choose the opposite route (using morphological exponence as a proxy for markedness) because determining the markedness of morphological exponence requires substantial additional research in morphophonology, which goes beyond the scope of our current project. Also, we submit that any progress here will have to look into degrees of overt exponence, along the lines suggested by Keine & Müller (2015).

that covers argument roles of both single-argument and two-argument predicates has a larger distribution, and is therefore unmarked in the classical sense of the term, than one that is limited to arguments of two-arguments predicates. As a general principle, then, we define the markedness of an alignment set in terms of whether or not the set contains an argument outside bivalent verbs, i.e. S. In the following we define markedness as follows, generalizing over all verb types:

(2) An alignment set α is marked relative to another alignment set β iff α contains argument roles from fewer numerical valence types than β, where the numerical valence types are monovalent, bivalent, and trivalent.

In the Vafsi example, this means that high-ranking P arguments are mapped into a marked alignment set (the {A,P} set), while low-ranking P arguments are mapped into an unmarked set (the {S,P} set), in line with Silverstein's predictions.

Under these assumptions, hypotheses of scale effects are specifically about marked vs. unmarked argument sets: we expect marked sets to associate preferentially with low-ranking A and high-ranking P arguments. If there is no difference in markedness, then all ranks on the scale show the same distribution, and there is no prediction. This is the case in the Vafsi past tense example with regard to NPs in A function: all nominal A arguments, regardless of their discourse status, are mapped into a marked alignment set, either {A} or {A,P}, and therefore always surface in the oblique case. By the same token, the hypotheses make no prediction on systems where arguments appear in different kinds of marked cases depending on their referential status – such as for example in Finnish, where some P arguments appear in the accusative while others appear in the partitive case. Since both cases define a marked alignment that contrasts P with {S,A}, there is no difference in markedness under the assumptions made in (2).

The predictions occasionally differ for A and P arguments, a difference enshrined in the traditional distinction between 'differential subject marking' and 'differential object marking'. Since in Vafsi all A arguments are marked, there is no prediction for A marking; for P arguments, by contrast, Vafsi is in line with the prediction that higher-ranking P arguments have a higher chance of being marked than lower-ranking P arguments. While in this case, there is a contrast between 'no prediction' and 'expected', some systems of alignment sets lead to conflicts in expectations. Khufi, another Iranian language, restricts the double-oblique system to a subset of pronouns (first and second person singular, third person) and contrasts this with neutral alignment in all other NPs. The

following data illustrate this: demonstrative (third person) pronouns are in the oblique case in A (3a) and P (3b) but not in S (3c) function; other pronouns and lexical nouns are always in the direct case (cf. the P arguments in 3a and 3d, the A argument in 3b and 3d and the S argument in 3e):

(3) Khufi (ISO639.3:sgh; Southeastern Iranian; Indo-European; Sokolova 1959)

a. *way* *xūðm* *wīnt.*
DIST.SG.OBL dream.DIR see.PST
A P
'He saw a dream.'

b. *mā́š=am* *way* *na* *talǽpt.*
1PL.DIR=1PL.PST DIST.SG.OBL NEG look.for.PST
A P
'We did not look for him.'

c. *yaw* *yat* *tar* *dum* *yīd.*
DIST.SG.DIR come.PST to MID.SG.OBL bridge.DIR
S
'He came towards that bridge.'

d. *Tarsakbṓy* *žær* *zṳ̊x̌t.*
Tarsakboy.DIR stone.DIR take.PST
A P
'Tarsakboy took the stone.'

e. *Tarsakbṓy* *xu* *ǰṓy-ti* *x̌āb* *na* *x̌ṳ̊vd.*
Tarsakboy.DIR REFL place=on night NEG sleep.PST
S
'Tarsakboy did not sleep at his place that night.'

Such a distribution is expected for P arguments: only high-ranking (first and second singular and all third person pronouns) P arguments are mapped into the marked {A,P} set; low-ranking P arguments are mapped into the unmarked {S,A,P} set. But for A arguments, the distribution is unexpected because high-ranking A arguments are also mapped into the marked set {A,P} while low-ranking arguments are mapped into the unmarked {S,A,P} set.

There are many possibilities of how markedness sets distribute across referential scales. Table 1 illustrates some of these with data we have in our database. In Table 1 we simply divide the scale into 'high', 'mid' and 'low', and spell out the concrete scales in the last column. Obviously, this begs the question of how referential scales are actually defined. We take this up in the following.

Table 1: A selection of observed distributions of case alignment sets across referential scales ('none' means 'no prediction', 'many' means 'predicted to be frequent', 'rare' means 'predicted to be rare or non-existent')

High	Mid	Low	Prediction for A	Prediction for P	Example	Relevant scale (segment) in example
{S,A}:{P}		{S,A,P}	none	many	Anamuxra (Ingram 2001)	N-anim > N-inanim
{S,A}:{P}	{S,P}:{A}		many	many	Dyirbal (Dixon 1972)	1/2 > 3/N
{S,A}:{P}	{S}:{A}:{P}	{S,P}:{A}	many	many	Djapu (Morphy 1983)	Pro > N-high-anim > other N
{S,A,P}		{S,A}:{P}	none	rare	Middle Atlas Berber (Pencheon 1973)	1/2/3 > N
{S,P}:{A}	{S}:{A}:{P}	{S,P}:{A}	none	rare	Gumbaynggir (Eades 1979)	3 > N-kin > N-other
{A,P}:{S}	{S,A,P}		rare	many	Khufi past tense (Sokolova 1959)	1s/2s/3 > 1p/2p/N
{A,P}:{S}	{S,A,P}	{A,P}:{S}	rare	rare	Vafsi past tense (Stilo 2004)	1p/1s > 2p > 2s/3p
{A,P}:{S}		{S,P}:{A}	none	many	Vafsi past tense (Stilo 2004)	N-high > N-low
{S}:{A}:{P}	{S,A}:{P}	{S}:{A}:{P}	rare	none	Talysh past tense (Schulze 2000)	1s > 2p/2s/3p > 3s
{S}:{A}:{P}		{S,P}:{A}	none	many	Nepali tense set I (Bickel 2011a)	anim/def > inanim/indef
{S}:{A}:{P}	{S,A}:{P}	{S}:{A}:{P}	many	many	Diyari (Austin 1981)	1s/2s > 1d/1p/2d/2p > 3

2.3 Defining referential scales

A referential scale is a scale defined by referential categories, covering inherent referential categories like 'animate', discourse-based referential categories like 'speaker' or 'proximative' and part of speech notions like 'pronoun'. Obviously, all these categories are ultimately language-specific and can only be identified by language-specific criteria (Haspelmath 2015). Yet, for many such categories, we can generalize over language-specific scales, because they show sufficient semantic overlap across languages. For example, it seems plausible that a category like 'first person singular' in one language is the same as the category 'first person singular' in another language. With categories like 'proximative' or 'topical', this is much less clear.

What is needed then is a list of category types that abstracts away from language-specific details and allows comparing language-specific referential categories, i.e. what is variously called 'typological types' (Bickel & Nichols 2002), 'values of typological features' (Haspelmath et al. 2005), or 'comparative concepts' (Haspelmath 2007, 2010). Notions like 'proximative', 'topical', 'definite' etc., for example, are probably best captured by a typological type like 'high discourse rank', which is defined in opposition to 'low discourse rank', with the understanding that 'discourse rank' is a probabilistic notion determined by a series of factors whose weights may differ from language to language.

Such type lists can be declared *a priori*, or they can be derived inductively by generalizing over all and only those language-specific categories that are encountered. Most lists that have been proposed in the literature are probably developed on the basis of a mix of *a priori* expectations and experience gained through typological survey work. Generally recognized types include notions like first, second, and third person; singular vs. dual vs. plural; pronoun vs. lexical noun; definite/topical vs. indefinite/nontopical; human vs. (nonhuman) animate vs. inanimate (e.g. Comrie 1981, Dixon 1994, Croft 1990). In our own database work we develop lists using the 'autotypologizing' method of Bickel & Nichols (2002): this method seeks to inductively abstract away from language-specific categories to exactly that degree that is needed to capture all language-specific distinctions encountered in a sample of language. In many cases this level of abstraction is fairly high, for example with notions like 'singular' or 'second person', which apply to a large number of languages, but in some cases, it is impossible to abstract away from an individual language. In our database, this can be illustrated with the arbitrary gender categories of German, which condition a case split, so that a distinct accusative case is limited to third person masculine pronouns and determiners and to first and second person pronouns.

Another type of split refers to discourse factors – a well-studied example of this is the factors determining object marking in English (Bresnan et al. 2007). While the factors are complex and include both language-specific and cross-linguistic categories, there is a general sense that the net effect of the factors is a broad distinction between higher vs. lower prominence in discourse, manifested variably as specific vs. nonspecific, definite vs. indefinite, topical vs. nontopical and similar such contrasts. We label this broad distinction by the term pair 'high' vs. 'low discourse rank', while noting that this glosses over substantial cross-linguistic variation (a point to which we will return in the Discussion section).

After surveying 435 languages with this method, we find the list of types in Table 2 to be at the right level of abstraction for capturing all distinctions ever made by case marking in at least one language. Language-specific categories which do not apply to more than one language have an arbitrary language ID number in their label, such as German (e.g. '3sgPro-masc87').

Given the list in Table 2, the question is how it maps into a scale. It has often been noted that the details of scales vary from language to language – e.g. some languages rank first person above second person while others rank second person above first person – but that there still are some basic principles – e.g. that all languages rank speech act participants above third persons. There are many proposals in the literature on what exactly these basic principles are, and in the following we explore an entire series of possible principles. In Section 4.2 we also compute a best-fitting scale empirically and explore this as well.

The hypothetical scales we test in the following are summarized in Table 3. For example, the 'SAP > 3/N' scale predicts that speech act participants rank higher than all other referents, but that languages can vary in the mutual ordering of first and second person and that differences in number are irrelevant, while the 'SAP > 3 > N' in addition predicts differential ranking between pronouns and nouns. The 'Pro > N' scale reduces this even further. The scale 'Pro/N-high > N-low' makes the cut slightly differently, capturing mainly effects from animacy, definiteness, specificity and related notions. The table lists two possible ranking of numbers. The sg > nsg ranking is based on the assumption that singular is more indexible than nonsingular and therefore ranks higher: singular items can be better pointed out than multiple items, in the same way as speech act participants can be better pointed at than other referents (Bickel & Nichols 2007). The reversed ranking 'nsg > sg' is based on the assumption that singular is structurally – and often also morphologically – unmarked relative to non-singular, and therefore ranks lower (Croft 1990).

Table 2: Referential categories referenced by case splits in the languages surveyed

Label	Definition
Pro	Pronouns. This refers to free pronouns that head NPs; it does not refer to pronominal agreement markers.
1sgPro	1st person singular pronoun
1duPro	1st person dual pronoun
1exclPro	1st person exclusive pronoun
1inclPro	1st person inclusive pronoun
1plPro	1st person plural pronoun
2sgPro	2nd person singular pronoun
2duPro	2nd person dual pronoun
2plPro	2nd person plural pronoun
3sgPro	3rd person singular pronoun
3duPro	3rd person dual pronoun
3plPro	3rd person plural pronoun
3Pro-anim	pronoun referring to an animate
3Pro-inanim	pronoun referring to inanimates
3sgPro-hum	3rd person singular pronoun with human reference
3sgPro-non-hum	3rd person singular pronoun with non-human reference
Pro-high	pronoun with a higher discourse rank than 'Pro-low' (where rank is determined by discourse factors with language-specific weights)
Pro-low	pronoun with a lower discourse rank than 'Pro-high' (where rank is determined by discourse factors with language-specific weights)
3Pro-high	3rd person pronoun (no number distinction) with a higher discourse rank than '3Pro-low' (where rank is determined by discourse factors with language-specific weights)
3Pro-low	3rd person pronoun (no number distinction) with a lower discourse rank than '3Pro-high' (where rank is determined by discourse factors with language-specific weights)
3sgPro-high	3rd person singular pronoun with a higher discourse rank than '3sgPro-low' (where rank is determined by discourse factors with language-specific weights)
3sgPro-low	3rd person plural pronoun with a lower discourse rank than '3sgPro-high' (where rank is determined by discourse factors with language-specific weights)
3plPro-high	3rd person plural pronoun with a higher discourse rank than '3plPro-low' (where rank is determined by discourse factors with language-specific weights)
3plPro-low	3rd person plural pronoun with a lower discourse rank than '3plPro-high' (where rank is determined by discourse factors with language-specific weights)
3sgPro-fem87	German 3rd person feminine pronoun
3sgPro-masc87	German 3rd person masculine pronoun
3sgPro-neut87	German 3rd person neutral pronoun

Table 2: *(continued)*

Label	Definition
N	lexical noun, nominalized verb – whether possessed or non-possessed (all)
N-anim	animate noun
N-inanim	inanimate noun
N-hum	human noun
N-non-hum-sg	non-human noun in singular
N-non-hum-du	non-human noun in dual
N-non-hum-pl	non-human noun in plural
N-proper	proper noun
N-common	common noun
N-common-sg	common (non-proper) noun in singular
N-common-pl	common (non-proper) noun in plural
N-def	definite noun
N-indef	indefinite noun
N-high	noun with a higher discourse rank than 'N-low' (where rank is determined by discourse factors with language-specific weights)
N-low	noun with a lower discourse rank than 'N-high' (where rank is determined by discourse factors with language-specific weights)
N-high-anim	noun denoting a higher animate (humans and some animals)
N-low-anim	noun denoting a lower animate (some animals)
N-spec	noun with specific reference
N-non-spec	noun without specific reference
N-kin	kin terms
N-non-kin	any noun apart from kin terms
N-non-kin-sg	any singular noun apart from kin terms
N-non-kin-pl	any plural noun apart from kin terms
N-pers	personal name (proper nouns which are personal names, but not toponyms, etc.)
N-pers-female	female personal name
N-pers-male	male personal name
N-non-pers	non-personal noun (common nouns and proper nouns which are not personal names (e.g. toponyms))
N-sg	noun in singular
N-pl	noun in plural
N-pl-anim	animate noun in plural
N-pl-inanim	inanimate noun in plural
N-masc-sg87	German masculine singular noun (case on determiner)
N-fem-sg87	German feminine singular noun
N-neut-sg87	German neutral singular noun
N1-sg-anim340	Russian animate noun of the inflectional class 1 (e.g. *student*)
N1-sg-inanim340	Russian inanimate noun of the inflectional class 1 (e.g. *zavod*, *mesto*)
N2-sg340	Russian noun of the inflectional class 2 (e.g. *komnata*, *muzhchina*, *sestra*)
N3-sg340	Russian noun of the inflectional class 3 (e.g. *doch'*, *noch'*)

2.4 Two models of scale effects

There are two models of how one can conceive of the way in which scales can determine the distribution of alignment sets. In the model that is traditionally assumed, scales predict a specific distribution of differential argument marking in grammatical systems: each grammatical system with a split either fits or does not fit the prediction, or, formulated as an implicational universal, 'if a language has a split in the case alignment of arguments, this split follows a universal scale'. We call this the 'Type Model'. The alternative, but so far largely unexplored model, is the 'Rank Model': in this model, scales are conceived of as ordered factors of categories that determine the relative probabilities of specific alignment sets for each category. In other words: the odds for case marking on a given argument correlate with the rank of that argument on a universal scale. In the following we discuss these models in more detail.

Type Models assess whether a split system of alignment sets in a given language fits vs. does not fit the predicted scale. For this, we define an alignment set as the set of argument roles selected by a case marker under a given referential condition (e.g. the set {S,A} selected by a nominative case under the referential condition 'third person')[5] and a system of alignment sets as the set of alignment sets defined by the case paradigm in a given language.

The criterion for fit is made explicit in (4) and relies on the same definition of markedness as in (2) above and the scales as defined in Section 2.3 ('higher' means to the left in the scales in Table 3 and 'position' refers to the set of categories between 'greater than' symbols in Table 3):

(4) For any given language, a system ℵ of alignment sets that shows one or more splits, fits a scale Ξ iff the categories mentioned in the definition of Ξ are part of the referential categories attested in ℵ, and
   a. for A arguments, no position on Ξ that contains a marked set containing A is ordered higher than a position that contains an unmarked set containing A.
   b. for P arguments, no position on Ξ that contains an unmarked set containing P is ordered higher than a position that contains a marked set containing P.

   A position Ξ_k contains a (un)marked set iff there exists a (un)marked alignment set that is defined for at least one category in Ξ_k.

5 Formally: given a set of roles R = {S,A,P} and a set of referential conditions C = {C1, . . . , C$_n$}, a case **K** in a given language (e.g. nominative case in Nepali) can be represented as $\mathbf{K} \subseteq R \times C$. The alignment set α of **K** for a given referential condition C_i is then $\alpha_{(\mathbf{K},C_i)} = \{R_i \in R \mid (R_i, C_i) \in \mathbf{K}\}$.

Table 3: *A priori* defined scales

Labels	Definition
1 > 2 > 3 > N	1sgPro/1duPro/1plPro > 2sgPro/2duPro/2plPro > 3sgPro/3plPro/3duPro/3sgPro-hum/3sgPro-non-hum/3sgPro-high/3plPro-high/3sgPro-low/ 3plPro-low/Pro-kin/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87 > N/N-hum/N-proper/N-anim/N-kin/N-def/N-indef/N-high-anim/N-low-anim/N-sg/N-pl/ N-spec/N-non-spec/N-inanim/N-non-kin/N-non-kin-sg/N-non-kin-pl/N-pers/N-non-hum-sg/ N-high/N-low/N-non-pers/N-common/N-non-hum-du/N-non-hum-pl/N-common-sg/ N-common-pl/N-pers-male/N-pers-female/N-masc-sg87/N-fem-sg87/N-neut-sg87/ N1-sg-inanim340/N1-sg-anim340/N2-sg340/N3-sg340/N-pl-anim/N-pl-inanim
SAP > 3/N	1sgPro/1duPro/1plPro/2sgPro/2duPro/2plPro > 3sgPro/3plPro/3duPro/3sgPro-hum/3sgPro-non-hum/N/N-hum/N-proper/N-anim/N-kin/ N-def/N-indef/N-high-anim/N-low-anim/N-sg/N-pl/N-spec/N-non-spec/N-inanim/N-non-kin/ N-non-kin-sg/N-non-kin-pl/N-pers/Pro-kin/N-non-hum-sg/N-high/N-low/N-non-pers/ N-common/N-non-hum- du/N-non-hum-pl/3sgPro-high/3plPro-high/3sgPro-low/3plPro-low/ 3sgPro-masc87/3sgPro-fem87/3sgPro-neut87/N-common-sg/N-common-pl/N-pers-male/ N-pers-female/N-masc-sg87/N-fem-sg87/N-neut-sg87/N1-sg-inanim340/N1-sg-anim340/ N2-sg340/N3-sg340/N-pl-anim/N-pl-inanim
SAP > 3 > N	2sgPro/1plPro/1sgPro/2plPro/1duPro/2duPro > 3sgPro/3plPro/Pro-kin/3duPro/3sgPro-hum/3sgPro-non-hum/3sgPro-high/3plPro-high/ 3sgPro-low/3plPro-low/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87 > N/N-hum/N-proper/N-anim/N-kin/N-def/N-indef/N-high-anim/N-low-anim/N-sg/N-pl/ N-spec/N-non-spec/N-inanim/N-non-kin/N-non-kin-sg/N-non-kin-pl/N-pers/N-non-hum-sg/ N-high/N-low/N-non-pers/N-common/N-non-hum-du/N-non-hum-pl/N-common-sg/ N-common-pl/N-pers-male/N-pers-female/N-masc-sg87/N-fem-sg87/N-neut-sg87/ N1-sg-inanim340/N1-sg-anim340/N2-sg340/N3-sg340/N-pl-anim/N-pl-inanim
SAP > 3 > N-high > N-low	1sgPro/1duPro/1plPro/2sgPro/2duPro/2plPro > 3sgPro/3plPro/3duPro/3sgPro-hum/3sgPro-non-hum/3sgPro-high/3plPro-high/3sgPro-low/ 3plPro-low/Pro-kin/3plPro-low/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87 > N-hum/N-proper/N-anim/N-kin/N-def/N-high-anim/N-spec/N-pers/N-high/N-pers-male/ N-pers-female/N1-sg-anim340/N-pl-anim > N-indef/N-low-anim/N-non-spec/N-inanim/N-non-kin/N-non-kin-sg/N-non-kin-pl/ N-non-hum-sg/N-non-hum-du/N-non-hum-pl/N-low/N-non-pers/N-common/N-common-sg/ N-common-pl/N1-sg-inanim340/N-pl-inanim
Pro > N	Pro/1sgPro/1duPro/1plPro/2sgPro/2duPro/2plPro/3sgPro/3plPro/3duPro/3sgPro-hum/ 3sgPro-non- hum/3sgPro-high/3plPro-high/3sgPro-low/3plPro-low/Pro-kin/3sgPro-masc87/ 3sgPro-fem87/3sgPro-neut87/Pro-high/Pro-low > N/N-hum/N-proper/N-anim/N-kin/N-def/N-indef/N-high-anim/N-low-anim/N-sg/N-pl/ N-spec/N-non-spec/N-inanim/N-non-kin/N-non-kin-sg/N-non-kin-pl/N-pers/N-non-hum-sg/ N-high/N-low/N-non-pers/N-common/N-non-hum-du/N-non-hum-pl/N-common-sg/ N-common-pl/N-pers-male/N-pers-female/N-masc-sg87/N-fem-sg87/N-neut-sg87/ N1-sg-inanim340/N1-sg-anim340/N2-sg340/N3-sg340/N-pl-anim/N-pl-inanim

Table 3: *(continued)*

Labels	Definition
Pro/ N-high > N-low	Pro/1sgPro/1duPro/1plPro/2sgPro/2duPro/2plPro/3sgPro/3plPro/3duPro/3sgPro-hum/ 3sgPro-non-hum/3sgPro-high/3plPro-high/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87/ Pro-high/Pro-low/Pro-kin/N-hum/N-proper/N-anim/N-kin/N-def/N-high-anim/N-spec/ N-pers/N-high/N-pers-male/N-pers-female/N1-sg-anim340/N-pl-anim > N-indef/N-low-anim/N-non-spec/N-inanim/N-non-kin/N-non-kin-sg/N-non-kin-pl/ N-non-kin-sg/N-non-kin-pl/N-non-hum-sg/N-non-hum-du/N-non-hum-pl/N-low/N-non-pers/ N-common/3sgPro-low/3plPro-low/N-common-sg/N-common-pl/N-pl-inanim/ N1-sg-inanim340
nsg > sg	1duPro/2duPro/3duPro/1plPro/N-pl/2plPro/3plPro/3plPro-high/3plPro-low/N-non-hum-du/ N-non-hum-pl/N-common-pl/N-pl-anim/N-pl-inanim/N-non-kin-pl > N-sg/2sgPro/3sgPro/1sgPro/N-non-hum-sg/3sgPro-hum/3sgPro-non-hum/3sgPro-high/ 3sgPro-low/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87/N-common-sg/N-masc-sg87/ N-fem-sg87/N-neut-sg87/N1-sg-inanim340/N1-sg-anim340/N2-sg340/N3-sg340/ N-non-kin-sg
sg > nsg	N-sg/2sgPro/3sgPro/1sgPro/N-non-hum-sg/3sgPro-hum/3sgPro-non-hum/3sgPro-high/ 3sgPro-low/N-non-hum-du/N-non-hum-pl/3sgPro-masc87/3sgPro-fem87/3sgPro-neut87/ N-common-sg/N-masc-sg87/N-fem-sg87/N-neut-sg87/N1-sg-inanim340/N1-sg-anim340/ N2-sg340/N3-sg340 > 1duPro/2duPro/3duPro/1plPro/N-pl/2plPro/3plPro/3plPro-high/3plPro-low/N-common-pl/ N-pl-anim/N-pl-inanim

Obviously, if a language does not reference any of the category types defined by a universal scale, e.g. if a language does not mark number as defined by the sg > nsg scale, the fit cannot be evaluated. In general, a language can be evaluated with regard to a scale Ξ only if each position of Ξ (as defined in Table 3), has a non-empty intersection with the category types referenced by the language.

Some of the possibilities defined by (4) can be illustrated by the patterns in Table 1 above. For example, Anamuxra case marking fits the N-anim > N-inanim scale for P arguments, but there is no prediction for the case marking of A arguments. Dyirbal fits the 1/2 > 3/N scale for both arguments. The past tense systems of Khufi, Vafsi or Talysh do not fit the respective scales in Table 1 with regard to the A argument, whereas Middle Atlas Berber, Gumbaynggir and the Vafsi past tense system do not fit with regard to the P argument. However, the scales in Table 1 are tailored to each language. The hypothesis that we aim to test is that there exists one or more *universal* scale(s) on which all systems fit, and the definition of fits in (4) targets these universal scales. For example, differential A marking in Diyari fits a number-related scale ranking nonsingular above singular referents. It also fits a SAP > 3 scale insofar as unmarked sets only occur

among speech-act participants, and so there is no case in which a marked set would ever outrank an unmarked set. But Diyari does not fit a person scale ranking first above second person because there are cases where a marked set (first person singular) outranks an unmarked set (second person plural). One could of course adjust the definition of the scale and condition – but then the scale is no longer universal.

The Rank Model is a standard logistic regression model: a scale is an ordered factor that is hypothesized to affect the probabilities of marked alignment sets. Specifically, the hypotheses to be tested are, for a given scale Ξ:

(5) a. For A: $\log\left(\frac{\pi(\text{marked})}{\pi(\text{unmarked})}\right) = \alpha - \beta\Xi$

b. For P: $\log\left(\frac{\pi(\text{marked})}{\pi(\text{unmarked})}\right) = \alpha + \beta\Xi$

That is, we hypothesize for A arguments, that the odds for marked alignment sets correlate negatively and significantly with Ξ, and for P arguments, that the odds for marked alignment sets correlates positively and significantly with Ξ.[6] This can be illustrated by the alignment systems in the Khufi participle-based tense (cf. (3) above): the difference between singular and plural speech act participants enters the analysis by different rank codings: for the sg > nsg scale, all singular pronouns are assigned rank 1, all dual and plural pronouns rank 2. The fact that all third person pronouns are (structurally) marked regardless of number is registered by the fact that they are all coded as having marked P arguments. For the regression model, speech act participants will increase the correlation between rank and markedness because only rank 1 is associated with marked P arguments; by contrast, third person P arguments will lower the correlation because they are marked on both rank 1 (singular) and rank 2 (nonsingular). If a language does not reference the categories of the relevant scale (e.g. makes no number distinction) or does not reference any category of the scale (because it has no split), it is irrelevant for the model.

The chief difference between the Type and the Rank Model is that the Rank Model is much less sensitive to individual exceptions. A scale may be a significant predictor overall even if, say, a specific pronoun, does not match the prediction. What counts is the overall trend. Under the Type Model, every single exception counts as evidence against a scale effect.

6 In the regression model we internally code 1 as the top (e.g. first person) and 1 + *n* (e.g. third person) as lower on the scale, in order to match standard (Western) linguistic parlance.

3 Testing universal effects: methods and data coding

3.1 Methods

As suggested by the preceding, Type Models can be evaluated by testing frequencies of fits against non-fits and Rank Models by logistic regression tests. However, simply applying such tests to raw data does not do justice to the hypothesis: all synchronic observations about language are the result of history, and therefore, any evaluation needs to target trends in diachrony rather than current distributions. If a scale is a genuine universal, we expect that each case system has a higher probability of developing in such a way as to conform to the scale than to contradict the scale. If the proto-language already fit the scale, we expect daughter languages, and therefore the whole family descending from the proto-language, to maintain the fit. If the proto-language did not fit the scale, we expect daughter languages to change their case systems so that they fit better. In either case, families will tend to be biased towards fitting the scale (i.e. for all families, there will be more members that fit than members that do not fit). By contrast, if the hypothesis of universal scale effects is wrong, we expect case systems to develop without regard to scales, subject to no or other principles. For example, languages could preserve case systems in pronouns and not in nouns (as they often do), regardless of whether this matches scales. As a result, families will be biased for or against a given scale at random.

In addition, it is well known (and has been emphasized since at least Dryer 1989) that typological distributions are not only affected by possible structural or cognitive principles – here, scales –, but also by areal diffusion resulting from language contact. In other words, the chances of finding a specific alignment set on a specific pronoun in a specific language may just as well be determined by the fact that the case distribution assimilated to neighboring languages, e.g. by calquing patterns of information structure. Therefore, any statistical test applied to typological data needs to control for the confounding factors of linguistic areas.

In the following we adopt Bickel's (2011b, 2013) Family Bias method of testing diachronic biases under area control. For each family we determine whether it is biased towards fitting a scale as opposed to not fitting the scale, using either the Type Model or the Rank Model. If the number of families with a bias towards the fit is significantly higher than the number of families not fitting the scale, and this is independent of area (as can be tested through loglinear modeling), the hypothesis is supported. Families can also be diverse (mixed), with no significant bias towards or against a given scale. As this can arise from both

imperfect developments towards a fit or away from a fit, diverse families provide no evidence on the hypothesis.

Family biases can be directly determined if families contain a sufficient number of members (in practice, at least $N = 5$). If there are fewer members or if we are dealing with isolates, we use the extrapolation strategies described in Bickel (2011b, 2013): if, say, 60% of large families are biased towards some specific structure (e.g. biased towards fitting a particular scale) rather than balanced between structures (i.e. with conflicting evidence), we estimate a .6 probability (as a 'prior' probability) that the members of small families come from larger unknown families with a bias as well (in whatever direction). Some of the known members will be representative of the bias in the unknown larger family, and so we can take their choice (e.g. towards fitting or towards not-fitting the scale) to reflect the bias. (For families with 2–4 members, we take the majority choice as reflecting the bias; if there is a tie, we pick one value at random.)

However, some known members will happen to be deviates, e.g. the odd guy(s) out that developed a non-fitting case system although the family as a whole is biased towards fitting the scale. The probability of being representative can be estimated from the strength of the bias in large families: e.g. if among biased large families, biases tend to be very strong (e.g. on average covering over 90% of members), we can estimate a high probability that the known members of small biased families are representative of the larger unknown family from which they derive; then, the probability of being the odd guy(s) out is much lower (though arguably never zero).

In summary, using the probabilities of bias and of representativity based on large families, we can estimate how many of the small families come from larger biased as opposed to unbiased families, and if they come from biased families, we can estimate whether the known members reflect the respective biases of their families or deviate from them. These extrapolation estimates introduce random error but do so along a normal distribution. Therefore, we get a fairly reliable estimate of family biases if we extrapolate many times (say one thousand times) and compute the average of this. All estimates of family biases in this paper are based on this procedure. For the taxonomy of families we rely on Nichols & Bickel (2009).

Note that none of the datasets we use is a random sample. Therefore, the principles of random-sampling theory are not applicable, and this makes it impossible to use statistical tests based on this theory. Following the suggestions by Janssen et al. (2006), we therefore employ exact and randomization tests, which test the probability of finding the observed distribution under reshuffling

of the data (as provided by the R Development Core Team (2012) for binomial tests and by Werft & Potter (2010) for logistic regressions).[7]

3.2 Data and data coding

Our database contains 435 languages. Most of these were surveyed by us, but about 20% of entries was taken from earlier work in the AUTOTYP project on typological databases (Nichols 1992, Bickel & Nichols 2009a,b). Since area and family factors can be best controlled if they are sampled densely, we specifically searched for areas and families with scale-based splits, collecting as many datapoints within these groups as we could.[8]

The database does not track alignment sets per se but instead codes each case in each language for the argument roles it covers, and if this coverage depends on the referential category of the argument role, the argument role is also specified for that category. For example, the database contains entries like 'Chantyal: ergative on A in all category types; nominative on S in all category types and on P in 'N-low'; accusative on P in 'N-high' and 'Pro'. From this, we compute the alignment sets for each referential category referenced by a language (via the set-theoretical derivation in footnote 5). In the Chantyal example, the alignment sets are {S,P} and {A} under the condition 'N-low', but {S}, {A}, {P} under all other referential conditions.

Apart from referential conditions, the set of arguments covered by a case marker is also sometimes conditioned by other factors, e.g. the distinction between inflectional forms (e.g. participle-based periphrastic forms vs. synthetic forms), between categories of verbs (e.g. realis vs. irrealis sets of forms) or between clause types (e.g. main vs. dependent or finite vs. nonfinite clauses). 27 languages in the database show one or more of these splits. For the sake of hypothesis testing, we enter these systems as independent datapoints into our computations in the same way as we enter two systems of genealogically related languages as independent datapoints. This raises the number of alignment systems to 462. Whether or not there are dependencies between subsystems within a language or between systems within related languages can then be statistically

7 An R function implementing the method is available at http://www.uzh.ch/spw/software. We use the function with its default parameter settings: most importantly, the threshold for taking a family to be biased is set at a significance level of $p < .1$ under an exact binomial or a permutation regression test and families count as 'large' if they contain at least 5 members. These choices are justified in Bickel (2013).

8 The database is available as an electronic appendix at http://www.spw.uzh.ch/autotyp/available.html

assessed by looking at family-internal distributions, i.e. by looking at developmental biases in the sense discussed before.

Another factor that can condition alignment sets is whether or not a particular argument is derived by diathesis, e.g. the German nominative case covers not only S and A, but also P arguments in passive constructions. We excluded all such derived roles because they are not relevant for the hypotheses as formulated above. By the same token, we exclude languages in which cases do not cover all referential categories because under some referential conditions, diathesis is obligatory and there is no possibility of expressing a given role without derivation (as is the case for instance in Yup'ik Eskimo where indefinite P arguments must be demoted to allative case via antipassivization; Reed et al. 1977, Mithun 1999). These cases evidence effects of referential categories, and possibly scales, but of a different kind than the ones under review in the present study (but see Bickel & Gaenszle 2007).

Yet another conditioning factor is lexical classes. Apart from the alignment patterns found with the majority of verbs (majority in the sense of lexical types and discourse tokens), many languages have minor alignment patterns limited to a subset of verbs, e.g. experiencer verbs, or verbs of excretion, or verbs of obligation and similar sets. An example is German, where some experiential predicates assign accusative to S arguments (e.g. *mich friert* 'I am cold'). For current purposes we leave such classes out of the picture and limit our attention to default (majority) classes because referential scale effects do not seem to interact with lexical classes so that scale structures vary across classes; all we observe is that effects get entirely blocked by specific non-default classes.

4 Results

In the following we first give an overview of the genealogical and geographical distribution of different A and P marking (Section 4.1). We then use our database to compute a scale of cross-linguistically recurrent referential categories based on their treatment by case markers (Section 4.2). Finally, we submit the resulting scales together with the hypothetical scales from Section 2.3 to tests under the assumption of a Type Model (Section 4.3) and of a Rank Model (Section 4.4).

4.1 Genealogical and geographical distribution

Of the 462 systems in the database, 59 have splits on A, i.e. differential A marking of any kind (fitting or not fitting scales), and 149 have splits on P, i.e. differential P marking of any kind; 41 systems have both splits at the same time. The

Table 4: A splits by family

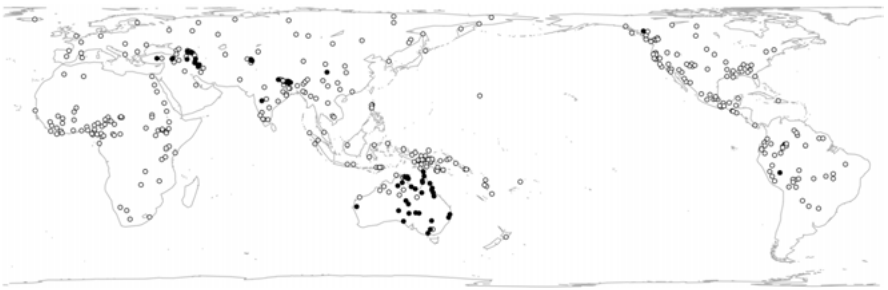
stock	*N*
Pama-Nyungan	29
Indo-European	15
Sino-Tibetan	8
Nakh-Daghestanian	3
Mangarayan	1
Pano-Tacanan	1
Tangkic	1
Tsimshianic	1

Table 5: P splits by family

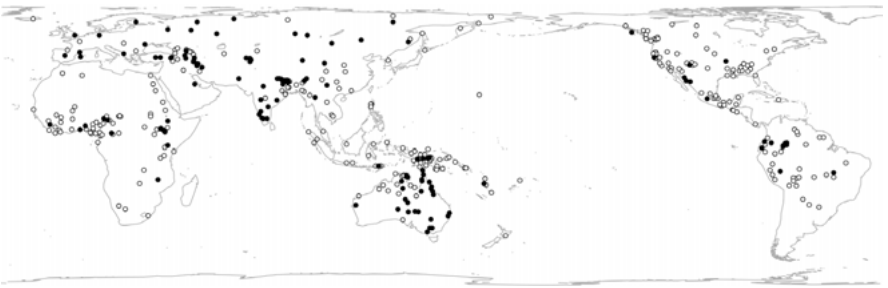
stock	*N*	stock	*N*	stock	*N*
Indo-European	38	Cushitic	2	Madang	1
Pama-Nyungan	26	Omotic	2	Mangarayan	1
Sino-Tibetan	13	Semitic	2	Mirndi	1
Dravidian	7	Adamawa-Ubangi	1	Nadahup	1
Turkic	7	Arawakan	1	Nakh-Daghestanian	1
Mongolian	4	Austroasiatic	1	Oksapmin	1
Tucánoan	4	Austronesian	1	Pano-Tacanan	1
Timor-Alor-Pantar	3	Awyu-Dumut	1	Pomoan	1
Uralic	3	Haida	1	Siouan	1
Uto-Aztecan	3	Kalam	1	South Atlantic	1
Barbacoan	2	Kusunda	1	Tarascan	1
Benue-Congo	2	Kwa	1	Tungusic	1
Chadic	2	Macro-Ge	1	Zuni	1

distribution of the splits across families, shown in Tables 4 and 5, is heavily skewed, restricted to 5.5% (8 out of 144) families in the database in the case of A splits and to 27% (39 out of 144) in the case of P marking. The top five families in the tables comprise 95% of all A splits and over 63% of all P splits.

The areal distribution of languages with splits is shown in Maps 1 and 2. In both types of splits, but especially in the case of differential P marking, there are frequency peaks in Eurasia (centered on Indo-Iranian languages, but deeply extending beyond this in the case of differential P marking; Bossong 1998) and in the New-Guinea/Australia – or 'Sahul' – macroarea (centered on Pama-Nyungan languages but extending to Tangkic and Southern New Guinea).



Map 1: Geographical distribution of languages with differential A marking of any kind (black dots) and languages without differential A marking (white dots)



Map 2: Geographical distribution of languages with differential P marking of any kind (black dots) and languages without differential P marking (white dots)

Both these macro-areas have been noted in previous work (e.g. Nichols 1992, 1993, 1997, Bickel & Nichols 2005, 2009b). We therefore tested their relative effects on the presence of splits. Like universal biases, areal biases are the result of diachronic trends. Therefore, if areas play a role, we expect them to affect the extent to which families are internally biased towards splits. In other words, if a family is located in a split-prone area, it is more likely to develop and/or maintain a split than outside the area. We test this using the Family Bias method described in Section 3.1 above.

A difficulty arises when families straddle area boundaries: Austronesian is split between languages in the larger Eurasian sphere (Southeast Asia) and the Sahul area (assuming the main boundary lying near the Wallace Line, as suggested by Nichols & Bickel 2009). Semitic is split between Eurasia (the Arabic peninsula and adjacent areas) and Africa (e.g. Amharic). In these two cases, we assess family biases within the areas separately, i.e. testing whether different

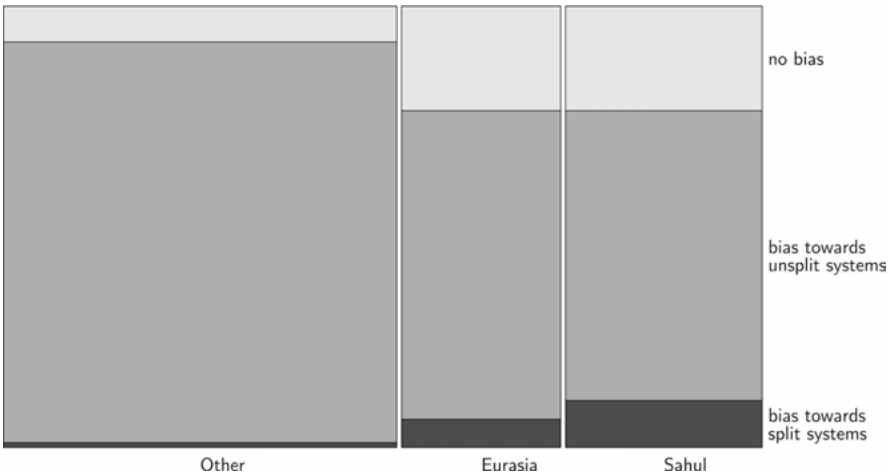


Figure 1: Estimated biases of families having split case marking for A across macro-areas. (The sizes of the individual tiles in the plot are proportional to frequencies, using the 'mosaic' plot technique provided by Meyer et al. 2006)

biases have developed depending on the area that part of the family is located in (see Bickel 2013 for general methodological discussion of this issue).

For the hypothesis of area-specific diachronic biases against or in favor of splits, diverse families are irrelevant since they can arise from imperfect biases towards *or* against splits. Thus, we tested whether the relative proportion of biases towards vs. against splits depends on area, using generalized linear modeling based on Poisson distributions (following Cysouw's (2010) suggestions; cf. Bickel 2011b). This is the case for A marking (Figure 1, likelihood ratio $\chi^2 = 6.17$, $p = .046$, $N = 135$ families). For P marking, there is no significant interaction between bias direction and areas (Figure 2, $\chi^2 = 3.95$, $p = .14$, $N = 120$ families), but this is largely due to the fact that Eurasia and the Sahul area have a similar proportions of families that are biased towards split. Taken together, the proportion in these areas is twice as high as in the rest of the world (estimated proportions: 31% split in Eurasia and Sahul vs. 14% split elsewhere, $\chi^2 = 4.62$, $p = .032$).

This suggests that splits as an abstract property are significantly affected by areal diffusion. Given this, it is imperative that any assessment of the precise nature of the splits – whether they fit a universal scale or not – control for areal diffusion.

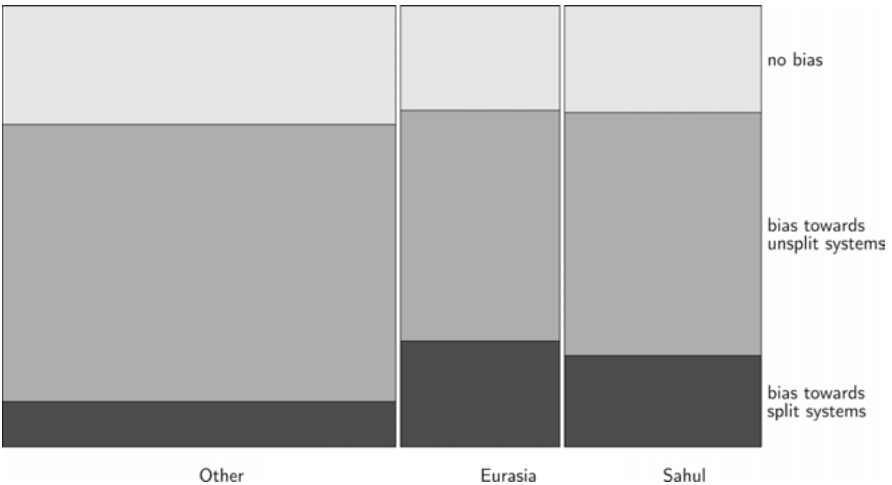


Figure 2: Estimated biases of families having split case marking for P across macro-areas (using the same mosaic plot techniques as in Figure 1)

4.2 Empirically derived scales

Scales can be thought of as summaries of similarity statements: A, B and C form a scale if both the pair of A and B and the pair of B and C are more similar to each other than A and C. As Cysouw (2015) points out, it is an empirical question to determine whether the similarity statements between such pairs of elements form a perfect one-dimensional scale, or whether they are not better represented by a multi-dimensional pattern of ordering: A, B, and C form a one-dimensional scale only to the extent that the similarity between A and C approximates the sum of the similarities between A and B and between B and C.

A general method for assessing the extent to which similarity patterns approximate a one-dimensional scale is what is called the Kruskal Stress (ϕ) in Multi-Dimensional Scaling. The key idea of Multi-Dimensional Scaling is to project a matrix of (dis)similarity statements onto a graph with k dimensions. The Kruskal Stress ϕ then measures the minimum extent to which the (dis)similarity statements have to be stretched or squeezed in order for all statements to be representable in k-dimensional space. For example, if A and C are equally similar to each other as each of the pairs {A, B} and {B, C}, projecting all three similarity statements onto a one-dimensional line incurs some amount of shrink-

Figure 3: The scale of split A marking as a one-dimensional solution to pairwise comparisons of whether referential categories of A arguments are treated the same way across languages (ϕ = 8.53%).

Figure 4: The scale of split P marking as a one-dimensional solution to pairwise comparisons of whether referential categories of P arguments are treated the same way across languages (ϕ = 9.79%).

ing the distances within {A, B} and within {B, C}, and stretching the distance between A and C.[9]

In order to empirically determine universal trends in the ordering patterns of referential categories, we selected those categories that are referenced by splits in at least two languages, separately for A and P. This results in 17 categories that condition A splits in 54 alignment systems, and 28 categories that condition P splits in 124 systems. From these tables, we then computed the dissimilarity between all pairs of categories by measuring the relative Hamming distance, which is the proportion of languages which split the pair of categories (so that one is in the marked the other in the unmarked set) among all languages that reference the pair of categories at all. The resulting matrix of pairwise similarity statements was then submitted to nonmetrical Multi-Dimensional Scaling (Kaufman & Rousseeuw 1990, Venables & Ripley 2002), projecting the matrix onto one dimension. The results are shown in Figures 3 and 4, together with the Kruskal Stress incured by the projection. For both A and P splits, the Kruskal Stress is reasonably low (ϕ < 10%), and inspecting higher-dimensional

9 Formally, $\phi = \sqrt{\frac{\Sigma_{i<j}(f(d_{i,j}) - D_{i,j})^2}{\Sigma_{i<j} D_{i,j}^2}}$, where $d_{i,j}$ are the observed and $D_{i,j}$ the projected dissimilarities, i.e. the formula basically computes the deviations of the projected from the observed distances, relative to the total of the projected distances. The function f is specific to what is called nonmetrical scaling and transforms distances so as to observe their rank order and abstracts away from their actual values. In the following we used nonmetrical scaling throughout.

solutions does not suggest further ordering patterns (nor does ϕ decrease much). Specifically and interestingly, no higher-dimensional solution points to, say, a dimension of number as opposed to a dimension of person.

For both A and P arguments, the solutions confirm what is called the Pro > N scale in Table 3, and this can be taken as an empirical confirmation of its cross-linguistic validity. In addition, the scales suggest particular rankings among persons and, for P marking only, between nouns that rank higher in discourse than other nouns:

(6) Empirically derived scales
   a. *A splits*
      1duPro > 2duPro > 1sgPro /1plPro /2sgPro > 3duPro /2plPro > 3sgPro /3plPro > N-kin/N-anim/N-high/N-def/N-indef/ N-non-kin/N-high-anim > N-low-anim/N-inanim
   b. *P splits*
      other pronouns > 1plPro/2plPro/3sgPro/3plPro > N-high > other N

The relatively low ranking of first person singular pronouns in the A scale matches our impressions: first person singular is indeed often treated differently from other persons. We briefly mentioned an example from the Australian language Arrernte above, and Bickel (2000) discusses the special status of first person singular A arguments in a number of Himalayan languages. For P splits, we have no immediate interpretation of the empirically derived scale, but we note that a high discourse status places NPs closer to pronouns than other nominal categories do (such as singular number, kinship, animacy etc.).

4.3 The Type Model

For all families containing languages with referential splits, we estimated biases towards or against fitting specific scales. As explained in Section 3.1, we performed these estimations by testing for biases in large families (via exact binomial tests) and extrapolating from this to small families and isolates. In order to control for the macro-areas (Eurasia and Sahul) that we observed to play a significant role in Section 4.1, we performed estimations separately for each macro-area.

Tables 6 and 7 show the number of families that are estimated to be biased towards ('+fit') or against ('–fit') case marking splits fitting a given scale within each macro-area.[10] Families estimated to be diverse (i.e. without evidence for a

10 Unlike in the geography report in Section 4.1, we display here absolute frequencies and not proportions because the absolute frequencies are much smaller.

Table 6: Estimated frequencies of families biased towards or against A-marking splits fitting a given scale, based on 144 families

Scale	Eurasia		Sahul		Other		*N*
	+fit	–fit	+fit	–fit	+fit	–fit	
1 > 2 > 3 > N	1.74	1.03	0	0	0	0	2.77
SAP > 3/N	1.49	0	0	0	0	0	1.49
SAP > 3 > N	1.51	0	0	0	0	0	1.51
SAP > 3 > N-high > N-low	0.32	0.01	0	0	0	0	0.33
Pro > N	1.51	0	2.29	0.1	0.52	0.47	4.89
Pro/N-high > N-low	1.56	0.1	1.62	0.05	0.02	0.5	3.86
nsg > sg	1.05	1.69	0	0	0	0	2.74
sg > nsg	0	1.48	0	1	0	0	2.48

Table 7: Estimated frequencies of families biased towards or against P-marking splits fitting a given scale, based on 144 families

Scale	Eurasia		Sahul		Other		*N*
	+fit	–fit	+fit	–fit	+fit	–fit	
1 > 2 > 3 > N	0.66	0.67	1.35	1.04	0.16	2.87	6.75
SAP > 3/N	0.78	0.53	1.21	1.12	1.23	2.19	7.06
SAP > 3 > N	0.66	0.69	1.32	1.04	0.35	2.58	6.63
SAP > 3 > N-high > N-low	0.34	0.01	0	0	0.03	0.49	0.87
Pro > N	12.89	1.92	5.93	0.39	8.15	2.75	32.04
Pro/N-high > N-low	8.11	0.08	2.8	0.18	4.55	0.49	16.21
nsg > sg	0	4.3	0.04	0.62	0.19	3.86	9
sg > nsg	2.38	1.98	0.66	1.7	2.23	1.78	10.73

diachronic bias) are not shown in the table since they do not give evidence for or against a hypothesized scale fit: diversity can arise through diachronic transitions both in line with and in contradiction to the hypothesis. This excludes most cases from outside Eurasia and Sahul. What is also omitted from the tables are results for the empirically derived scales in (6). These scales do not suggest any family biases at all. The reason for this is that the empirically derived scales average across the details of specific languages while the Type Model is defined with respect to language-specific ways of fitting vs. not fitting a scale. (The empirically derived scales are better testable in the Rank Model; see Section 4.4 below.)

The results suggest that there is no consistent and area-independent effect for a specific scale to match A and P marking at the same time. For A-marking

splits, the number of families estimated to be biased towards fits is in the same ballpark as the number of families estimated to be biased against fits. In each case, the overall numbers are very small because only few languages and few families reference the necessary categories for evaluation under a Type Model. As a result, even when there are biases, these biases are limited to one or two families, and typically only in Eurasia: almost all evidence is limited to Sino-Tibetan and/or Indo-European. The strongest evidence in Table 6 comes from the part-of-speech scale Pro > N and the discourse-based Pro/N-high > N-low scales, but even here the total number of families estimated to be biased towards fitting the scale is below two or three and usually involves only Indo-European in Eurasia and Pama-Nyungan in Sahul.

For P-marking splits, the part-of-speech and the discourse-based scales do seem to show area-independent effects. Table 7 suggests that for these two scales, families estimated to be biased towards fitting the scales outnumber families estimated to be biased against fitting the scales, and this holds in all three macro-areas. However, the overall count is still small (estimated at N = 32.04 for the part-of-speech scale and at N = 16.21 for the discourse-based scales), and the differences in counts is appreciably strong only in Eurasia (with approximately 13:2 and 8:0 ratios). The low overall counts of families with relevant splits also makes the extrapolation procedure problematic because the procedure can rely on only very few large families – in fact only 4 in Eurasia (Indo-European, Dravidian, Sino-Tibetan and Turkic) and 1 (Pama-Nyungan) in Sahul. This leaves much guess work in the extrapolation procedure,[11] and even just a handful of further families could alter the results substantially. The small number of data and the uncertainty that comes with this precludes an overall test of whether the higher counts of families fitting the two scales is statistically significant.

4.4 The Rank Model

Estimating family biases under a Rank Model follows basically the same procedure as under a Type Model except for the way biases are tested in large families: under the rank model, a bias towards a scale means, for A-marking, that the scale is a significant predictor in a logistic regression of case-marking and that

11 The prior probabilities for families to be biased (in any direction) tend to have large 95% credibility intervals in this dataset: [.48, .99] in Eurasia, [.16, .99] in Sahul and a whopping [0,1] elsewhere. (These intervals indicate the ballpark of where our estimates can be placed on the basis of the analysis of large families.)

the predictor has a negative coefficient ($-\beta\Xi$ in 5), i.e. the odds for marked alignments (such as ergative alignment) decrease when going up the scale. A bias against the scale means that the scale is again a significant predictor but that its coefficient has a reversed sign ($+\beta\Xi$). The same evaluations of biases hold for P marking, but with all signs reversed: a bias towards a given scale means that the scale is a significant predictor with a positive coefficient ($+\beta\Xi$); a bias against the scale will have a negative coefficient ($-\beta\Xi$). While the logic of these tests is straightforward, the small number of datapoints per family poses a problem for regression models. Some of these problems can be resolved by relying on permutation tests for the regressions (Werft & Potter 2010), but even then regressions can fail to converge. We interpret non-convergence and other failures in the regressions as lack of evidence for the kind of straightforward regression that we would expect if there had been a systematic diachronic bias aligning case distributions with a scale.

An additional complication arises for the extrapolation procedure of the Family Bias method. As detailed in Section 3.1, the extrapolation procedure combines statistical information from large families with observations from small families: if the distribution across large families lets us expect a small family or isolate to reflect a bias, and we estimate its member(s) to be representative of this bias, we take the direction of the bias to be given by the observed value (if there is only one member; if there are more, we take the majority value, picking one at random in the case of ties). Under a Rank Model, the analogue of observed values in small families cannot itself be a logistic regression (which is a statistical model and not a single observation). Instead, we use the observation on scale 'fits' that we used in the Type Model and take these as 'pseudo-regression' data-points when assessing the direction (positive or negative) of the regression in the small families that we estimate to be biased. Concretely: for A marking, a Type-Model fit, i.e. a fit between markedness distributions and a given scale as defined in (4a), corresponds to a negative pseudo-regression (and thus in favor of the hypothesis). Absence of a fit, i.e. a distribution that violates a given scale, counts as a positive pseudo-regression, contradicting the hypothesis. Finally, if there is conflicting evidence (e.g. because the markedness distribution does not reference all relevant categories of a given scale), but we still estimate the small families to be biased (because of the estimated prior bias probability), we choose the direction of the pseudo-correlation at random (assuming that the diachronic bias could be in any direction). The same logic applies to P marking, with all signs reversed.

Tables 8 and 9 show the results, again excluding the many families estimated to be diverse, i.e. without evidence for a diachronic bias. As noted in Section 2.4, the Rank Model is less sensitive to individual exceptions than the

Table 8: Estimated frequencies of families biased towards A-marking odds correlating negatively (–) or positively (+) with a given scale, based on 144 families

Scale	Eurasia		Sahul		Other		*N*
	–	+	–	+	–	+	
1 > 2 > 3 > N	4.60	3.60	7.60	6.80	1.40	1.40	25.40
SAP > 3/N	6	7	7.40	5.20	1.20	1.60	28.40
SAP > 3 > N	4.60	3.60	7.60	6.80	1.40	1.40	25.40
SAP > 3 > N-high > N-low	6.40	3	6.80	5.40	1.80	1	24.40
Pro > N	3.40	2.20	8	5.80	1.80	1.20	22.40
Pro/N-high > N-low	1	1	5.80	4.40	1.40	1.20	14.80
nsg > sg	2.80	2.60	8.80	4.40	0.80	1.20	20.60
sg > nsg	1.80	3.60	7.80	5.40	0.80	1.20	20.60
Empirical Scale (6a)	5.29	3.39	7.38	6.35	1.44	1.44	25.30

Table 9: Estimated frequencies of families biased towards P-marking odds correlating positively (+) or negatively (–) with a given scale, based on 144 families

Scale	Eurasia		Sahul		Other		*N*
	+	–	+	–	+	–	
1 > 2 > 3 > N	3.21	2.24	7.28	6.26	2.84	3.96	25.79
SAP > 3/N	3.21	2.24	7.28	6.26	2.84	3.96	25.79
SAP > 3 > N	5.36	3.43	7.29	6.25	2.98	3.91	29.22
SAP > 3 > N-high > N-low	9.69	5.75	10.53	10.6	5.41	5.26	47.23
Pro > N	7.63	4.36	12.6	8.31	4.17	2.8	39.87
Pro/N-high > N-low	12.13	6.52	7.74	6	6.66	4.01	43.05
nsg > sg	2.28	3.26	3.12	3.15	4.35	6.33	22.5
sg > nsg	3.28	2.26	3.12	3.15	6.35	4.33	22.5
Empirical Scale (6b)	9.60	5.66	7.30	6.35	6.28	4.34	39.53

Type Model and can therefore pick up overall trends much better. This results in much higher overall bias estimations (ranging from 14.8 to 47.23 families, with a mean of 27.37), giving more robust results.

For A marking (Table 8), the hypothesis is that families develop biases so that the odds for marked alignments (ergativity) correlate negatively with a universal scale. Our results suggest that families are in fact just as likely to develop negative or positive regressions, with differences staying below an estimated 2.4 families and mean of .56 families. Stronger differences are limited to a single macro-area: there is a slight trend of the complex SAP > 3 > N-high > N-low scale to reveal more families biased towards negative than towards positive

regressions (with about a 6:3 ratio), but the effect is limited to Eurasia and therefore not universal. In the Sahul macro-area, there is a trend (with about a 9:4 ratio) for families to be biased so that alignments correlate with a nsg > sg scale, but no such effect is replicated elsewhere. Interestingly, not even the empirical scale (6a) shows a clear effect: families estimated to be biased towards this scale outnumber the opposite bias only by 1.9 in Eurasia, by 1.03 in Sahul and not at all elsewhere.

The results for P marking (Table 9) reveal a similar picture, with most differences staying below 2 families and averaging at .25. The empirically derived scale (6b) fares slightly better but the difference falls short of statistical significance (under a Poisson model controlling for area effects: $\chi^2 = 1.26$, $p = .26$). Similar to what we observe under the Type Model, more remarkable trends emerge for the part-of-speech (Pro > N) and the discourse-based scale (Pro/N-high > N-low). However, the differences are again relatively small and not statistically significant ($\chi^2 = 2.53$, $p = .12$). The discourse-based scale shows a strong difference only in Eurasia (with about a 12:7 ratio). The same is true for the more complex scale SAP > 3 > N-high > N-low (ca. 10:6 ratio) that also has – as noted above – an appreciable effect on A marking in Eurasia only.

In the preceding we observed several cases where scale effects show up only in specific macro-areas. However, none of the relevant differences reach statistical significance – except for the Pro > N scale and P marking, where the difference between Sahul and the rest of the world is significant ($\chi^2 = 6.20$, $p = .013$).

5 Discussion

Regardless of how one spells out the hypothesis of universal scale effects on case marking, our results show that there tend to be just as many families with diachronic biases in support of the hypothesis as there are families with diachronic biases against the hypothesis. If there are appreciable differences in frequencies they are limited to just one macro-area. These results are direct evidence against universal scale effects.

A possible exception from this is constituted by the part-of-speech (Pro > N) and the discourse-based (Pro/N-high > N-low) scales, which reveal area-independent effects under the Type Model for P marking. However, as noted in the Results section, under the Type Model, the evidence for this is based on very small numbers of families and is therefore not conclusive. Under the Rank Model, the evidence for the part-of-speech scale falls short of statistical significance, and it is limited to Eurasia in the case of the discourse-based scale.

This means that for these two scales we lack evidence against a universal effect although we also lack solid evidence in favor of universal effects. Interestingly, the two scales are those that fit least the spirit of the overall idea of scale effects: the part-of-speech scale could just as well be interpreted as a simple pronoun vs. noun distinction that has in fact nothing to do with any scale or hierarchy. The discourse-based scale can also be conceived of as a binary distinction of discourse-prominent vs. other referents. Moreover, as noted in Section 2.3, such a distinction lacks a reliable cross-linguistic interpretation because its constitutive categories ('high' vs. 'low') vary widely from language to language. Thus, even if we were to discover more families with diachronic biases in favor of these scales, it is doubtful whether they can be taken as support for genuinely universal and genuinely scalar effects.

We also investigated the possibility of deriving scales from the bottom up (Section 4.2). The resulting scales average over the way case marking systems are distributed over referential categories. As such they are ill–fitted for the Type Model which evaluates fits separately for each case system in each language and is highly sensitive to individual exceptions. But the empirically derived scales are suitable for the Rank Model where family biases are assessed on the basis of overall trends in regression models. Remarkably, however, the empirically derived scales did not reveal clear universal trends and the number of families that tend to violate them are not much lower than the number of families that support them (see the last rows in Tables 8 and 9).

While we find evidence against universal scale effects or, in two cases, no evidence in favor of such effects, our study reveals strong areal effects: families tend to develop referentially-conditioned alignment splits (of whatever kind, scalar or not) significantly more often in the Eurasia and Sahul macro-areas than anywhere else (cf. Section 4.1). In one case, the area difference is shown by a specific scale: for P arguments, the odds for marked alignments depend on a Pronoun > Noun scale significantly more often in the Sahul macro-area than anywhere else. Beyond this, area differences are not sensitive to specific scales but to an overall split in the abstract.

6 Conclusions

The impression of universal scale effects in the literature seems to stem from the ubiquity of such effects in Eurasia and Sahul. As soon as one controls for possible effects of areal diffusion, the numbers of families that can be statistically estimated to have developed in line with the hypothesis are in the same ballpark

as the number of families that can be estimated to have developed in contradiction to the hypothesis. The only scale effects where the evidence is more ambiguous than this are effects of parts-of-speech and discourse rank in P marking. However, even if the evidence became less ambiguous, this would not necessarily speak for a universal hypothesis of scale effects on case alignments because the relevant distinction are fundamentally binary and not scalar in nature.

The present study confirms the results of Bickel & Witzlack-Makarevich (2008), which uses a smaller dataset (353 vs. 462 case systems in the present study) and a different approach to the statistical testing of linguistic universals. In the 2008 study we tested the Type Model only within sufficiently large families and assessed the Rank Model by modeling family and area membership as parameters of a single regression model. The present study relies on statistical estimates of diachronic biases. Because this can be done across all families, regardless of their size, the method is able to pick up more distributional signals in the data. As a result, we can now strengthen our claim: with the possible exceptions noted, we have now evidence for the absence of universal effects, while the 2008 study only suggested absence of evidence for universal effects.

Another finding emerging from the present study is that differential case marking on A and P is first and foremost a pattern prone to diffusion. However, what seems to diffuse is splits in the abstract and not splits tied to specific scales (with the possible exception of the pronoun vs. noun distinction in Sahul). The details of the splits seem to vary strongly across languages and are subject to idiosyncratic developments of the kind discussed by Garrett (1990) and Filimonova (2005): reanalyses of individual case markers or case attrition in nouns as opposed to pronouns. Given these findings, what becomes an urgent task now is research into the ways in which splits spread in language contact. We submit that any deeper understanding of referential scale effects in individual languages needs to explore how it arose diachronically and what role was played in this by area diffusion – especially in the Eurasia and Sahul macro-areas.

References

Aissen, Judith. 1999. Markedness and subject choice in Optimality Theory. *Natural Language and Linguistic Theory* 17. 673–711.

Austin, Peter K. 1981. *A grammar of Diyari, South Australia*. Cambridge: Cambridge University Press.

Bickel, Balthasar. 2000. Person and evidence in Himalayan languages. *Linguistics of the Tibeto-Burman Area* 23. 1–12.

Bickel, Balthasar. 2008. On the scope of the referential hierarchy in the typology of grammatical relations. In Greville G. Corbett & Michael Noonan (eds.), *Case and grammatical relations: papers in honor of Bernard Comrie*, 191–210. Amsterdam: Benjamins.

Bickel, Balthasar. 2011a. Grammatical relations typology. In Jae Jung Song (ed.), *The Oxford handbook of language typology*, 399–444. Oxford: Oxford University Press.

Bickel, Balthasar. 2011b. Statistical modeling of language universals. *Linguistic Typology* 15. 401–414.

Bickel, Balthasar. 2013. Distributional biases in language families. In Balthasar Bickel, Lenore A. Grenoble, David A. Peterson & Alan Timberlake (eds.), *Language typology and historical contingency*, 415–444. Amsterdam: Benjamins.

Bickel, Balthasar & Martin Gaenszle. 2007. Generics as first person undergoers and the political history of the Southern Kirant. Paper presented at the 7th Biannual Meeting of the Association for Linguistic Typology, Paris, September 26, 2007, http://www.uzh.ch/spw/bickel/presentations/Kiranti1U_ALT2007.pdf.

Bickel, Balthasar & Johanna Nichols. 2002. Autotypologizing databases and their use in fieldwork. In Peter Austin, Helen Dry & Peter Wittenburg (eds.), *Proceedings of the International LREC Workshop on Resources and Tools in Field Linguistics, Las Palmas, 26–27 May 2002*, Nijmegen: MPI for Psycholinguistics [http://www.uzh.ch/spw/autotyp/download/canary.pdf].

Bickel, Balthasar & Johanna Nichols. 2005. Areal patterns in the World Atlas of Language Structures. Paper presented at the 6th Biannual Conference of the Association for Linguistic Typology, Padang, July 24; available at http://www.uzh.ch/spw/autotyp/download.

Bickel, Balthasar & Johanna Nichols. 2007. Inflectional morphology. In Timothy Shopen (ed.), *Language typology and syntactic description*, 169–240. Cambridge: Cambridge University Press (Revised second edition).

Bickel, Balthasar & Johanna Nichols. 2009a. Case marking and alignment. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford handbook of case*, 304–321. Oxford: Oxford University Press.

Bickel, Balthasar & Johanna Nichols. 2009b. The geography of case. In Andrej Malchukov & Andrew Spencer (eds.), *The Oxford handbook of case*, 479–493. Oxford: Oxford University Press.

Bickel, Balthasar & Alena Witzlack-Makarevich. 2008. Referential scales and case alignment: reviewing the typological evidence. In Andrej Malchukov & Marc Richards (eds.), *Scales (Linguistische ArbeitsBerichte 86)*, 1–37. Leipzig: Institut für Linguistik [http://www.uni-leipzig.de/~asw/lab/lab86/LAB86_Bickel_Witzlack.pdf].

Bickel, Balthasar, Alena Witzlack-Makarevich, Taras Zakharko & Giorgio Iemmolo. in press. Exploring diachronic universals of agreement: alignment patterns and zero marking across person categories. In Jürg Fleischer, Elisabeth Rieken & Paul Widmer (eds.), *Agreement from a diachronic perspective*, Berlin: de Gruyter.

Bossong, Georg. 1998. Le marquage différentiel de l'objet dans les langues de l'Europe. In Jack Feuillet (ed.), *Actance et valence dans les langues de l'Europe*, 193–258. Berlin: Mouton de Gruyter.

Bresnan, Joan, Anna Cueni, Tatiana Nikitina & R. Harald Baayen. 2007. Predicting the dative alternation. In G. Bouma, I. Kraemer & J. Zwarts (eds.), *Cognitive foundations of interpretation*, 69–94. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

Comrie, Bernard. 1981. *Language universals and linguistic typology*. Chicago: University of Chicago Press.

Croft, William. 1990. *Typology and universals*. Cambridge: Cambridge University Press.

Cysouw, Michael. 2010. On the probability distribution of typological frequencies. In Christian Ebert, Gerhard Jäger & Jens Michaelis (eds.), *The Mathematics of Language*, 29–35. Berlin: Springer.

Cysouw, Michael. 2011. Understanding transition probabilities. *Linguistic Typology* 15. 415–431.
Cysouw, Michael. 2015. Generalizing scales. This volume.
DeLancey, Scott. 1981. An interpretation of split ergativity and related patterns. *Language* 57. 626–657.
Dixon, R. M. W. 1972. *The Dyirbal language of North Queensland*. Cambridge: Cambridge University Press.
Dixon, R. M. W. 1994. *Ergativity*. Cambridge: Cambridge University Press.
Dowty, David R. 1991. Thematic proto-roles and argument selection. *Language* 67. 547–619.
Dryer, Matthew S. 1989. Large linguistic areas and language sampling. *Studies in Language* 13. 257–292.
Dryer, Matthew S. 1992. The Greenbergian word order correlations. *Language* 68. 81–138.
Eades, Diana. 1979. Gumbainggir. In R. M. W. Dixon & Barry J. Blake (eds.), *Handbook of Australian Languages 1*, 244–361. Amsterdam: Benjamins.
Filimonova, Elena. 2005. The noun phrase hierarchy and relational marking: problems and counterevidence. *Linguistic Typology* 9. 77–113.
Garrett, Andrew. 1990. The origin of NP split ergativity. *Language* 66. 261–296.
Haig, Geoffrey L. J. 2008. *Alignment change in Iranian languages: a construction grammar approach*. New York: Mouton de Gruyter.
Haspelmath, Martin. 2007. Pre-established categories don't exist: consequences for language description and typology. *Linguistic Typology* 11. 119–132.
Haspelmath, Martin. 2009. An empirical test of the Agglutination Hypothesis. In Sergio Scalise, Elisabetta Magni & Antonietta Bisetto (eds.), *Universals of language today*, 13–29. Berlin: Springer.
Haspelmath, Martin. 2010. Comparative concepts and descriptive categories: consequences for language description and typology. *Language* 86. 663–687.
Haspelmath, Martin. 2015. Descriptive scales versus comparative scales. This volume.
Haspelmath, Martin, Matthew S. Dryer, David Gil & Bernard Comrie (eds.). 2005. *The world atlas of language structures*. Oxford: Oxford University Press.
Ingram, Andrew. 2001. *A grammar of Anamuxra: a Language of Madang Province, Papua New Guinea*: University of Sydney dissertation.
Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga. 2006. Randomization tests in language typology. *Linguistic Typology* 10. 419–440.
Kaufman, Leonard & Peter J. Rousseeuw. 1990. *Finding groups in data: an introduction to cluster analysis*. New York: Wiley.
Keine, Stefan & Gereon Müller. 2015. Differential argument encoding by impoverishment. This volume.
Kiparsky, Paul. 2008. Universals constrain change; change results in typological generalizations. In Jeff Good (ed.), *Linguistic universals and language change*, 23–53. Oxford: Oxford University Press.
Merlan, Francesca. 1982. *Mangarayi*. Amsterdam: North-Holland.
Meyer, David, Achim Zeileis & Kurt Hornik. 2006. The strucplot framework: visualizing multiway contingency tables with vcd. *Journal of Statistical Software* 17. 1–48.
Mithun, Marianne. 1999. *The languages of Native North America*. Cambridge: Cambridge University Press.
Moravcsik, Edith. 1978. On the distribution of ergative and accusative patterns. *Lingua* 45. 233–279.

Morphy, Frances. 1983. Djapu, a Yolngu dialect. In Robert M.W. Dixon & Barry J. Blake (eds.), *Handbook of Australian Languages* 3, 1–188. Amsterdam: Benjamins.

Nichols, Johanna. 1992. *Linguistic diversity in space and time*. Chicago: The University of Chicago Press.

Nichols, Johanna. 1993. Ergativity and linguistic geography. *Australian Journal of Linguistics* 13. 39–89.

Nichols, Johanna. 1997. Sprung from two common sources: Sahul as a linguistic area. In Patrick McConvell (ed.), *Archeology and linguistics: global perspectives on Ancient Australia*, Melbourne.

Nichols, Johanna & Balthasar Bickel. 2009. The AUTOTYP genealogy and geography database: 2009 release. Electronic database, http://www.uzh.ch/spw/autotyp.

Pencheon, Thomas G. 1973. *Tamazight of the Ayt Ndhir*. Los Angeles: Undena.

R Development Core Team. 2012. *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing, http://www.r-project.org.

Reed, Irene, Osakito Miyako, Steven Jacobsen, Paschal Afcan & Michael Krauss. 1977. *Yup'ik Eskimo grammar*. Fairbanks: Alaska Native Language Center, University of Alaska.

Schulze, Wolfgang. 2000. *Northern Talysh*. Munich: Lincom Europa.

Silverstein, Michael. 1976. Hierarchy of features and ergativity. In R. M. W. Dixon (ed.), *Grammatical categories in Australian languages*, 112–171. New Jersey: Humanities Press.

Sokolova, Valentina Stepanovna. 1959. *Rusanskie i chufskie teksty i slovar'*. Moskva: Nauka.

Stilo, Donald. 2004. *Vafsi Folk Tales*. Wiesbaden: Reichert.

Venables, W. N. & Brian D. Ripley. 2002. *Modern applied statistics with S*. New York: Springer.

Werft, Wiebke & Douglas M. Potter. 2010. glmperm: Inference in Generalized Linear Models. R package version 1.0-3, http://CRAN.R-project.org/package=glmperm.

Wilkins, David. 1989. *Mparntwe Arrernte (Aranda): studies in the structure and semantics of grammar*: Australian National University dissertation.